PB84-177971

Data Management for Quality Assurance

Research Triangle Inst.
Research Triangle Park, NC

Prepared for

Industrial Environmental Research Lab.
Research Triangle Park, NC

Apr 84

U.S. Department of Commerce
National Technical Information Service
NTIS

PB84-177971

EPA-600/D-84-103
April 1984

DATA MANAGEMENT FOR QUALITY ASSURANCE

by

Clinton E. Tatsch and Donald E. Lentzen
Research Triangle Institute
Research Triangle Park, NC 27709

and

Gary L. Johnson
Industrial Environmental Research Laboratory
Research Triangle Park, NC 27711

EPA Contract
68-02-3146

EPA Project Officer
Gary L. Johnson

INDUSTRIAL ENVIRONMENTAL RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NC 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/D-84-103	2.
3. RECIPIENT'S ACCESSION NO. PB84 177971	
4. TITLE AND SUBTITLE Data Management for Quality Assurance	5. REPORT DATE 1984
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Clinton E. Tatsch and Donald E. Lentzen (RTI), and Gary L. Johnson (EPA, IERL-RTP)	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Research Triangle Institute P. O. Box 12194 Research Triangle Park, North Carolina 27711	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO. 68-02-3146
12. SPONSORING AGENCY NAME AND ADDRESS EPA, Office of Research and Development Industrial Environmental Research Laboratory Research Triangle Park, NC 27711	13. TYPE OF REPORT AND PERIOD COVERED Journal Article; February 83
	14. SPONSORING AGENCY CODE EPA/600/13
15. SUPPLEMENTARY NOTES	

16. ABSTRACT The article describes the application of quality assurance (QA) principles to two programs supporting EPA's Environmental Assessment Data System (EADS). The ultimate objective of any measurement program is the creation of data that is processed to provide information. The development of information (e.g., trends, parameter estimates) from data is a process that may be as difficult as the data production, if not more so. Application of QA principles to this interpretive phase of information production includes: (1) data acquisition (human/machine interface), (2) verification (range- and relational-checking), (3) data processing (transfer, transformations, and reduction), and (4) presentation of results (comparability, ease of interpretation).

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Pollution	Pollution Control	13B
Quality Assurance	Stationary Sources	13H, 14D
Measurement	Environmental Assessment	14G
Data Acquisition		14B
Verifying		
Data Processing		09B

18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS (This Report) Unclassified	21. NO. OF PAGES
	20. SECURITY CLASS (This page) Unclassified	22. PRICE

EPA Form 2220-1 (9-73)

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

DATA MANAGEMENT FOR QUALITY ASSURANCE

CLINTON E. TATSCH AND DONALD E. LENTZEN
RESEARCH TRIANGLE INSTITUTE
POST OFFICE BOX 12194
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709
AND
GARY L. JOHNSON
U. S. ENVIRONMENTAL PROTECTION AGENCY
INDUSTRIAL ENVIRONMENTAL RESEARCH LABORATORY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

ABSTRACT

The ultimate objective of any measurement program is the creation of data which is processed to provide information. The development of information (e.g., trends, parameter estimates) from data is a process that may be as difficult as the data production, if not more so. Application of quality assurance (QA) principles to this interpretive phase of information production includes:

- data acquisition (human-machine interface),
- verification (range- and relational-checking),
- data processing (transfer, transformations, and reduction), and
- presentation of results (comparability, ease of interpretation).

Application of such principles to two programs supporting EPA's Environmental Assessment Data System (EADS) is described.

GENERAL TEXT

INTRODUCTION

The first definition of "data" by Webster's Dictionary [1] is "...factual information used as a basis for reasoning, discussion, or calculation." It is the "...reasoning, discussion, or calculation..." which is of ultimate interest. The data are useful only to the extent that they accurately and readily promote the reasoning. An obvious corollary is that the reasoning process can be only as good as the data upon which it stands. If the data are inaccurate, it is certain that the reasoning can be no better.

It is within this context that the U.S. Environmental Protection Agency began implementation of an Agency-wide Quality Assurance program in 1979, encompassing "all environmentally related measurements"[2]. Progress in the development of this program is continuing and applies to all analytes and sample matrices of interest to EPA. Collation of data quality measures is in progress; evaluating the data quality of ongoing projects and increasing the awareness of personnel to data quality issues are also part of this program.

Another area of importance in the link between the data and the reasoning process, is that the data not only accurately represent a phenomenon, but that they do it readily. This aspect of data processing is important, since the more readily accurate data in their various forms are understood, the more likely they are to be properly applied in the reasoning process. However, this challenging aspect involves human perception and the need to anticipate the immediate needs of the data users, as well as those who use the data long after they are collected. For the purpose of this discussion, the term "information" will be used to refer to data which have been summarized and are presentable in a readily understandable form. Thus "information" might refer to graphs, diagrams, or tables as examples of possible representations.

The process of obtaining data from which to construct information is what is normally addressed in discussions of data quality. Such issues as experimental design, sampling errors, and analytical errors are in the usual domain of consideration. For the rest of this discussion, it will be assumed that these have been suitably addressed, and the focus of attention will be on the aspects of delivering accurate and readily understandable information to users. Also discussed is how this may be used to enhance the quality of the gathered data by providing appropriate feedback to data gatherers.

THE MODEL

A model for viewing data, after it has been created, is shown in Figure 1. This model is valid for all types of data - whether they be burette readings in a wet chemistry laboratory or output from computerized, instrumented data acquisition systems of the most modern design.

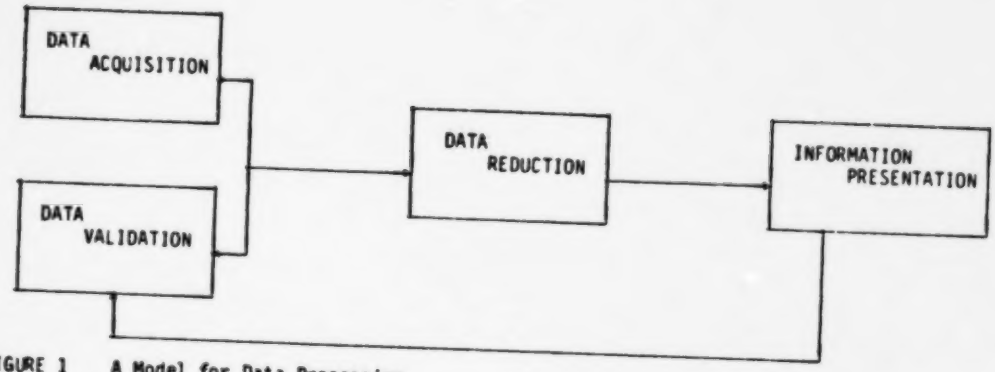

FIGURE 1 A Model for Data Processing.

Data Acquisition involves recording each datum in a useable form which can be copied. This involves some form of permanent media and a means of writing: examples are laboratory notebooks and magnetic tapes.

Data Validation is the step which often is performed subconsciously during manual data collection. It may involve a priori notions of data characteristics, such as the range of reasonable values. It may also involve a posteriori checks of data relationships for which there is a historical precedent, such as sampling volume or filter loadings in air measurement studies. Such checks serve as indications of data invalidity, in a tentative sense. Both of these steps apply to the observed data, and only rarely to calculated or derived information.

Data Reduction is the step in which data are summarized in the context of a physical, chemical, or statistical model. In general, it is this step which begins the irreversible production of information: one can always reconstruct the summary from the data, but the data cannot be reconstructed from the summary. This step requires careful scientific judgement as to appropriateness of the model to the specific situation. It also requires careful execution in the implementation of that judgement in the mundane, step-by-step protocol which associates the proper data with the much fewer generated results.

Following what is frequently preliminary data reduction, the intermediate results must be combined in some form of **Information Presentation** which further summarizes the intermediate results to produce information. This information may be in the form of tables designed to accentuate similarities or contrasts in results. It may be organized as a series of graphs depicting trends and other relationships. While several common formats are in use for presenting information, there remains the growing challenge to adequately summarize masses of data in a comprehensible manner. The difficulty remains as to just what the reader or viewer can readily comprehend. This problem is being intensified by the increasing amounts of data which must be digested. It is also being alleviated by the availability of inexpensive computing power and the developing graphics techniques for producing information directly from data. Application of computing power to two different but related types of data quality problems will be discussed below. One involves processing data from a large environmental data base; the second involves interactive processing of data for the same data base.

Both approaches to information production are related to the Environmental Assessment Data Systems (EADS). The EADS is a comprehensive system of computerized data bases that describes multimedia discharges from energy systems and industrial processes. It consists of four major data bases: the Fine Particle Emissions Information Systems (FPEIS), the Gaseous Emissions Data System (GEDS), the Liquid Effluent Data System (LEDS), and the Solids Discharge Data System (SDDS). The FPEIS is a repository for particle size distribution measurements. All of EPA's Industrial Environmental Research Laboratory's particle sizing data are stored in EAPS/FPEIS. It is installed at EPA's National Computer Center (NCC) at Research Triangle Park, NC, on the UNIVAC U-1100/83 computer. The EADS structure recognizes the sensitivity of some multimedia discharge data. If such data is labeled "confidential" by the submitter, the site name or location will remain anonymous.

Data for EADS/FPEIS originate from two major sources. The first is particle size data which have been collected, and are submitted after a project has been completed. Another source is data which are in the process of being collected, that may be utilized during the testing program.

PADRE-I

The first example of the application of quality assurance to data analysis is a program for processing particle size distributions, the data for which have previously been loaded into the FPEIS data base. These data are accessible through language interfaces which permit processing in FORTRAN and data base accesses through subroutine calls. Thus, the appropriate data elements from the data base are available for use by the program which has been named PADRE (for PArticle Data REduction). The impetus for PADRE has consistently been the aspect of results presentation, in terms of information which can be readily accessed by data base users. An important aspect of quality assurance in software design is that the objective be clearly specified at the outset of the project. In this research environment, it is expected that the desires and needs of the users will expand as they realize what capabilities are available in terms of information presentation.

To describe the approach to data management in PADRE, the format outlined in Figure 1 will be followed. The Data Acquisition aspect of the process has been completed prior to PADRE's operation; data have been encoded, loaded, and given a preliminary screening as part of routine data base function. Data Validation is then one of the first steps in PADRE.

The first type of data validation procedure is simple range-checking. There are criteria specific to each data item resulting from considering not only what data are possible, but their effect on subsequent processing steps. For example, since sampling time and flow rates in impactor trainscan never be negative, such values are not reasonable. Additionally, there are criteria for the finally produced information which must necessarily override the data values. For example, the data must be extrapolated to at least 30 micrometers in order to permit estimation of particle size distributions in this region: DMAX (the observer-estimated maximum particle diameter) is frequently below 30 micrometers and thus 30 micrometers is used during processing rather than DMAX. A warning is printed in the regular program output as to the alteration, and the data summary is flagged to alert PADRE users to the type of action taken.

The next action of PADRE is Data Reduction. Selected data are combined according to standard procedures used in the Cascade Impactor Data Reduction System (CIDRS)[3] to produce points on a diameter mass concentration distribution. For some purposes, this also serves as adequate Information Presentation. However, it was decided that interpolation at selected standard diameters should be included in PADRE. Thus, concentration values can subsequently be processed using the standard EADS/FPEIS access language. Appropriate curve-fitting algorithms and codes, available in the procedures noted above [3], were adapted to this application. While there is the capability of plotting either cumulative or derivative size-mass concentration distributions, only the fitted values at preselected diameters are extracted and stored in the data base.

An advantage of this program is that it has been possible to evaluate data in ways which would normally be impossible. For example, it will not be obvious if low values of sample flowrate have been erroneously entered until the results are available and one notes that the concentration distribution is unrealistically large. This is just one example of how aspects of results presentation can be used as a means for checking the quality of the data from which it was derived. This aspect of a feedback loop for improving the quality of information was a major consideration in developing the program which will be described next.

PADRE-II

An aspect of data quality which receives little formal attention relates to the feedback of useful information to the persons who are gathering the data. In many manual measurement systems, such as cascade impactor measurements for obtaining particle-size distribution information, it is not unusual for the information to become available only weeks after the conclusion of data collection. This obviously precludes adjustments of a test plan in response to test conditions. It also prevents the test crew from evaluating the quality of their own efforts. This program provides this information to test crews, in the field, at their test site, in a fashion which would encourage them to increase the quality of their data. In order to appreciate the operational problems and their sensitivity to external influences, it is necessary to briefly describe an impactor test.

Use of a cascade impactor involves initial weighings of a series of impactor substrates on which dust particles of various sizes will impact. These weighings for use in stationary source testing must typically be accurate to at least 0.1 milligram. These substrates are then carefully placed on the respective impactor stages as the impactor is assembled. Following leak-checking and warmup of the assembled impactor to at least the flue gas temperature from which it will be sampling, sample of the dusty gas is extracted from the gas stream for a specified length of time using a standard sampling train. The impactor is designed so that particles of successively smaller diameters impact on the substrates as the sample stream progresses through the impactor. The diameter at each stage may be calculated based on parameters such as gas temperature, density, and velocity at the stage. At the end of the sampling run, the impactor is carefully removed from the gas stream, cooled, and disassembled. The substrates on which the particles have impacted are

reweighed, and from the knowledge of total gas volume through the impactor, the concentration of the particles in the gas stream can be inferred for each substrate and its respective size particles. In this way, one obtains the particle-size versus mass-concentration distribution of the sampled gas stream, as soon as the calculation can be completed.

The inital aspects of this project were reasonably well-defined. The observables were such parameters as gas temperature, pressure, and composition, impactor calibration constants, and stage weight changes. The information initially expected to be useful was similar to that produced by PADRE-I; namely, various aspects of the particle-size/mass-concentration distribution. Since this was to be an interactive program, user inputs were to be in response to prompts from the program, using test crew conventions. The output was to be in a concise, yet ample, format. Response time was to be less than a couple of seconds, and no observable data should need entry into the computer more than once. Additional requirements were that user security be appropriate to protect each user's data from unauthorized alteration or inspection, as well as to allow addition and editing of observed data at any time and from any location where a portable computer terminal could be used. The system which implements these requirements is called PADRE-II, since it is the second in a series of related programs.

One of the principal motives for developing PADRE-II was to obtain accurate observed data, in order to produce valid information. Hence the interactive data acquisition phase involves a significant amount of preliminary data validation in various levels of complexity, with user feedback as soon as possible. First, data are entered into the PADRE-II system as responses to queries from the system. This ensures that users do not accidentally overlook the necessary data items. Responses are checked for possible typographical errors as they are received by the program. For example, if characters are received when numeric data are expected (for items such as flowrate or time of day), the user is informed immediately, and asked to re-enter the value for the requested datum. Once a datum has been validated at this level, it is stored in the data file. Once all data have been entered and stored, a report is printed at the computer terminal showing all the entered data and permitting revision of values by the user. When no further editing is requested, PADRE-II continues to the next stage of processing. This describes all phases of data entry for the impactor test, which includes substrate pretest weighing, test conditions, and substrate post-test weighings which generally occur at different times and locations.

When all required data have been validated, users are asked then, and only then, if they wish to begin data reduction. The reduction is based on the logic of CIDRS [3], and thus is subject to the verification of data reduction against other impactor data reduction systems. It also includes a more sophisticated level of data validation. For example, research indicates that stage jet velocities in excess of certain values are likely to induce particle bounce, while jet velocities below certain other values may lead to erroneously calculated particle diameters. When such conditions are encountered, PADRE-II gives appropriate warnings in its printed output as part of the observable data validation process.

The information presentation is in the form of a standard report which lists stage loadings and calculated diameters, along with such information as the observed (and cumulative) particle mass concentrations. Such information is useful to test crews who need to obtain data for a specified particle size range, since impactor operations do not lend themselves to intuitive estimation of the particle size captured on each stage. Calculations and outputs are suitable for the final report, as well.

As in PADRE-I, it was anticipated that users would compare results from different runs, such as those at the inlet and the outlet of a control device. For this reason, the curve-fitting capability of PADRE-I is also included. As this phase of the project was completed, it was believed that PADRE-II would supply suitable information in a form which was readily usable to test crews. They could rapidly and reliably reduce their impactor data for field (and laboratory) use with minimal user training.

As PADRE-II went through its testing phase and into production use, requests for additional information presentation capabilities began to accumulate. In response to these requests, PADRE now features statistical analysis of results from multiple runs as well as the capability for producing initial data plots at the terminal. Control device efficiency calculations, comparing inlet and outlet concentrations, have also been included. These capabilities now permit users to reduce data in the field, interspersed with their data collection activities. This permits rapid adjustment to actual conditions of an impactor test, increasing the quality of the data as the result of having accurate and readily understandable information available.

SUMMARY

In summary, quality assurance of measurements extends beyond data generation activities such as sampling and analysis. It also includes the conversion of observable data into usable information. Aspects of this conversion process include:

- data acquisition,
- data validation,
- data reduction, and
- information presentation.

Two different examples are described which illustrate the utility of this process in providing information as a means of enhancing data quality. Where possible, it is desirable to have accurate and easily understandable information rapidly available to personnel who generate data, in order to increase the accuracy of the subsequent data. Development of understandable information is challenging, and is strongly subject to the needs of the recipient of the information.

REFERENCES

1. Woolf, H. B. (Editor), Webster's New Collegiate Dictionary. G. & C. Merriam Company, Springfield, MA, 1975, p. 286.

2. Quality Assurance Management Staff, U. S. E.P.A., "Strategy for the Implementation of the Environmental Protection Agency's Mandatory Quality Assurance Program," QAMS-001/80, March 11, 1980.

3. Larkin, R. J. (Editor), "Environmental Assessment Data Systems: Systems Overview Manual," EPA-600/8-80-005 (NTIS PB 80-218266), January 1980.

4. Johnson, J. W., G. I. Clinard, L. G. Felix, and J. D. McCain, "A Computer-based Cascade Impactor Data Reduction System," EPA-600/7-78-042 (NTIS PB 285-433), March 1978.